POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Gabriele Ginestroni**
Giacomo Gumiero
Lorenzo Iovine
Nicola Landini
Francesco Leone

Group Number: **10**

Academic Year: 2022-2023

Contents

1 | Introduction

In this chapter will be presented the problem specification and the hypothesis under which the database is implemented.

1.1. Problem Specification

This project aims to build an Information System that handles scientific articles contained in the DBLP bibliography. The project involves managing the type of the articles and the associated DOI (Digital Object Identifier), which identifies an article or a document and links to it on the web. Other entities to deal with are authors, identified by an ID or ORCID (Open Researcher and Contributor ID), and their affiliations with organizations. In order to address the problem, we will store data in a graph database, allowing us to visualize relations and handle information correctly.

1.2. Assumptions

1. All the data in the dataset are heterogeneous, so fields are different
2. The **authors** with missing field *_id* are not considered
3. It is possible that an author writes for different organizations
4. Field *_id* in **author** is unique
5. Field *_id* in **article** exists and it is unique
6. It is impossible that 2 different articles are on the same journal, in the same *volume* with an intersection between *page_start* and *page_end*
7. The designed model doesn't take into consideration the URL associated to the article node, as the main focus of the project was not reading the article
8. It is possible to find a self-reference in a publication
9. A *venue* can be instantiated as a journal, a conference or a generic venue!!!!!!!!

2 | ER Diagram

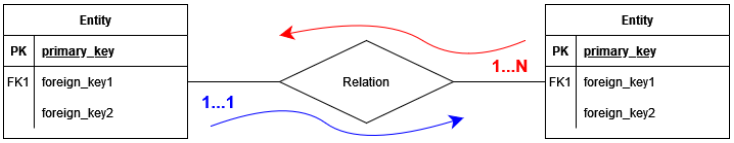

Figure 2.1: ER Diagram Organization

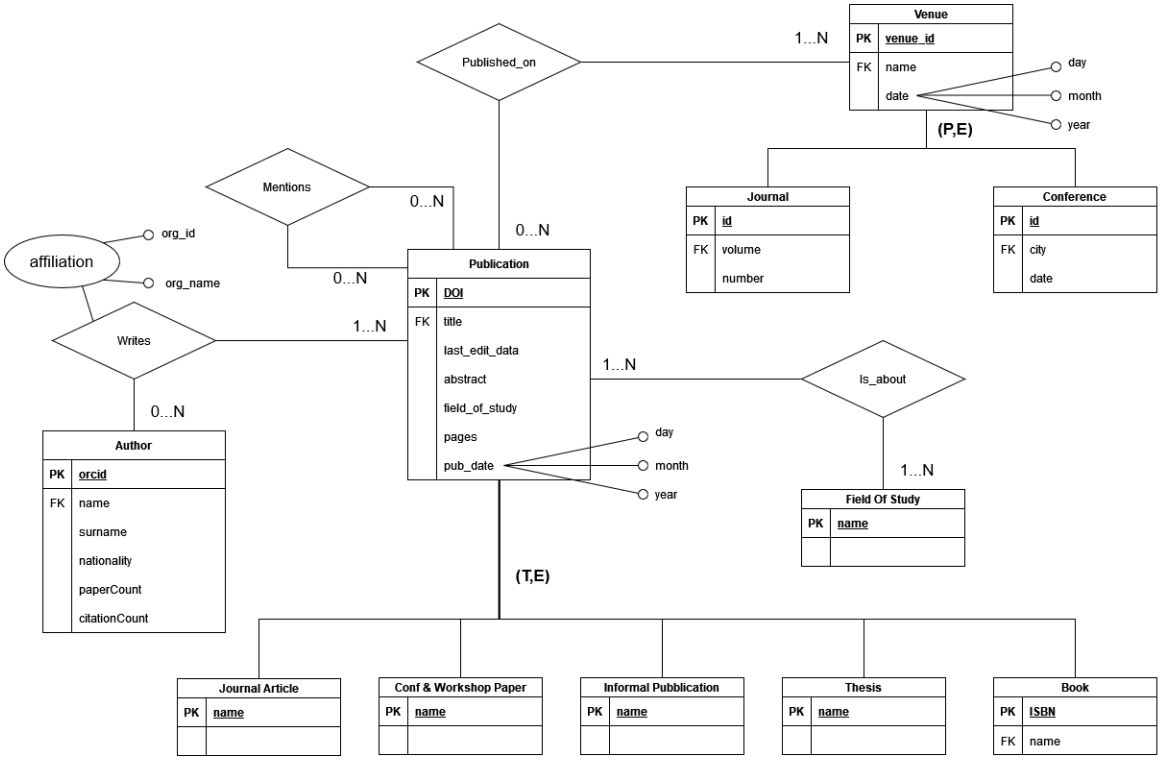

Figure 2.2: ER Diagram

The ER diagram designed contains the following entities:

- **Publication:** this entity represents all the scientific articles. They are identified by their primary key *_id* and other important attributes are: *DOI, title, last_edit_data, abstract, keywords, pages, pub_date*. Of course the attributes of such entity could be enlarged, but as a sample dataset we have believed these are enough. Publication entity is the superclass of a Total and Exclusive ISA relationship with the following subclasses: *Journal article, Conference & Workshop paper, Informal publication, Thesis, Book*
- **Author:** it represents all the people that submitted at least one publication. Its primary key is *_id* and the foreign keys are: *name, surname, nationality, paperCount, citationCount*. Of course the attributes of such entity could be enlarged, but as a sample dataset we have believed these are enough
- **Venue:** it's the entity that represents the type of a publication. This is a superclass that creates a Partial and Exclusive ISA relationship with the two subclasses *Journal* and *Conference*. The primary key is *raw* and the other keys are: *name, date, venue_id*
- **Field Of Study:** this entity represents the topics of the related publication

The ER diagram designed contains the following relationships:

- **Writes:** is the relationship between *author* and *publication* which specifies also the affiliation with the org_id and the org_name
- **Mentions:** occurs between two *publication* and specifies when a publication refers to another one
- **Is about:** binds a *publication* with its *fields of study*
- **Published on:** simply relates a *publication* to its *venue*

3 | Dataset Description

4 | Graph Diagram

5 | Queries and Commands

6 | Conclusion

A | Appendix A

If you need to include an appendix to support the research in your thesis, you can place it at the end of the manuscript. An appendix contains supplementary material (figures, tables, data, codes, mathematical proofs, surveys, ...) which supplement the main results contained in the previous chapters.

List of Figures

List of Tables